PROBABILISTIC REASONING OVER TIME

Outline

- Time and uncertainty
- Inference in temporal models
- Hidden Markov models
- Dynamic Bayesian networks

Based on the textbook by Stuart Russell & Peter Norvig:

Artificial Intelligence, A Modern Approach (2nd Edition)

Chapter 15; excluding Sections 15.4 and 15.6

1. TIME AND UNCERTAINTY

- We have previously developed our techniques for probabilistic reasoning in the context of **static** worlds.
- E.g. when repairing a car, it is assumed that whatever is broken remains broken during the process of diagnosis.
- However, in certain domains **dynamic** aspects become essential.

Example. A doctor is treating a diabetic patient.

— Recent insulin doses, food intake, blood sugar measurements, and other physical signs serve as pieces of evidence.

— The doctor decides about food intake and insulin dose.

States and Observations

- The process of change is viewed as a series of snapshots, each of which describes the state of the world at a particular time.
- Each **time slice** involves a set of random variables indexed by t:
  1. the set of *unobservable* state variables $\mathbf{X}_t$ and
  2. the set of *observable* evidence variables $\mathbf{E}_t$.
- The observation at time t is $\mathbf{E}_t = \mathbf{e}_t$ for some set of values $\mathbf{e}_t$.
- The notation $\mathbf{X}_{a:b}$ denotes the sets of variables from $\mathbf{X}_a$ to $\mathbf{X}_b$.
- The interval between time slices depends on the problem!

Stationary Processes and the Markov Assumption

- In a **stationary process**, the changes in the world state are governed by laws that do not themselves change over time.
- A **first-order** Markov process satisfies an equation

$$\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1})$$

  where $\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1})$ forms the **transition model** of the process.
- In addition, it is typical to assume a **sensor model** of the form

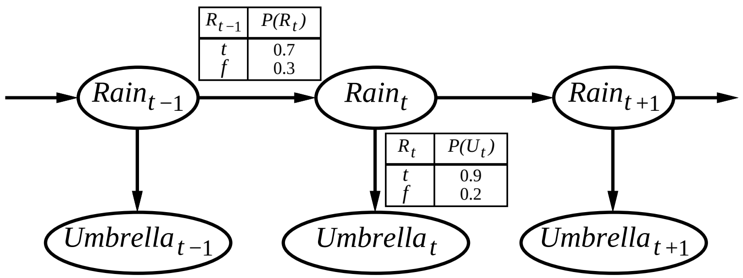
$$\mathbf{P}(\mathbf{E}_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{1:t-1}) = \mathbf{P}(\mathbf{E}_t \mid \mathbf{X}_t)$$

  so that observations depend on the current state only.

Example. A security guard is working at some secret underground installation and would like to know whether it is raining today or not.

The only access to the outside world occurs each morning when the director comes in with, or without, an umbrella.

- The set of state variables $\mathbf{X}_t = \{Rain_t\}$ for $t = 0, 1, \ldots$.
- The set of evidence variables $\mathbf{E}_t = \{Umbrella_t\}$ for $t = 1, 2, \ldots$.



Resulting Joint Distribution

- In addition to transition and sensor models, we need to specify a prior distribution $\mathbf{P}(\mathbf{X}_0)$ over the state at time 0.
- Combining this with the preceding transition and sensor models, which are *independence assumptions*, implies a distribution

  $$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \mathbf{P}(\mathbf{X}_0) \prod_{i=1}^{t} \mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1}) \mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i).$$

  for any point of time t.
- If necessary, the Markov assumption can be recovered by introducing suitable state variables.

Example. When modelling a battery-powered robot wandering in the *xy*-plane, the battery level has to be taken into account.

Example

Recall the security guard perceiving umbrellas.

The probability of an event $\neg r_0, \neg r_1, \neg r_2, u_1, u_2$ is calculated as follows:

1. The prior distribution $\mathbf{P}(R_0) = \langle 0.5, 0.5 \rangle$.
2. Next, we apply transition and sensor models over all time steps:

$$\begin{aligned} & P(\neg r_0, \neg r_1, \neg r_2, u_1, u_2) \\ = \; & P(\neg r_0) \times P(\neg r_1 \mid \neg r_0) P(u_1 \mid \neg r_1) \times P(\neg r_2 \mid \neg r_1) P(u_2 \mid \neg r_2) \\ = \; & 0.5 \times 0.7 \times 0.2 \times 0.7 \times 0.2 \\ = \; & 9.8 \times 10^{-3}. \end{aligned}$$

2. INFERENCE IN TEMPORAL MODELS

Having set up the generic temporal model, we may formulate the basic inference tasks that are to be solved.

1. **Filtering** or **monitoring**: the task is to compute the *belief state*, i.e. the posterior distribution $\mathbf{P}(\mathbf{X}_t \mid \mathbf{e}_{1:t})$ over the current state.
2. **Prediction**: the posterior distribution $\mathbf{P}(\mathbf{X}_{t+k} \mid \mathbf{e}_{1:t})$ over the *future* state is of interest for some $k > 0$.
3. **Smoothing** or **hindsight**: the aim is to compute $\mathbf{P}(\mathbf{X}_k \mid \mathbf{e}_{1:t})$ where $0 \leq k < t$ for some *past* state.
4. **Most likely explanation** is a sequence of states $\mathbf{x}_{1:t}$ that maximizes $P(\mathbf{x}_{1:t} \mid \mathbf{e}_{1:t})$ for the observations $\mathbf{e}_{1:t}$ to date.

Filtering

- In **recursive estimation**, the idea is to compute $\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1})$ as a function of $\mathbf{e}_{t+1}$ and $\mathbf{P}(\mathbf{X}_t \mid \mathbf{e}_{1:t})$.
- Using transition and sensor models we obtain by Bayes' rule, conditioning, and the Markov assumption that

$$\begin{aligned}\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) &= \alpha\mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}, \mathbf{e}_{1:t})\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{e}_{1:t}) \\ &= \alpha\mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1})\textstyle\sum_{\mathbf{x}_t}\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{x}_t)P(\mathbf{x}_t \mid \mathbf{e}_{1:t}).\end{aligned}$$

- This can be viewed as the propagation of a message $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t \mid \mathbf{e}_{1:t})$ *forward*: $\mathbf{f}_{1:t+1} = \alpha\,\text{FORWARD}(\mathbf{f}_{1:t}, \mathbf{e}_{t+1})$.
- The time and space requirements for updating are constant!

Example

The security guard has a prior belief $\mathbf{P}(R_0) = \langle 0.5, 0.5\rangle$ about the state.

1. The prediction from $t = 0$ to $t = 1$ gives

$$\begin{aligned}\mathbf{P}(R_1) &= \textstyle\sum_{r_0}\mathbf{P}(R_1 \mid r_0)P(r_0) = \mathbf{P}(R_1 \mid r_0)P(r_0) + \mathbf{P}(R_1 \mid \neg r_0)P(\neg r_0) \\ &= \langle 0.7, 0.3\rangle \times 0.5 + \langle 0.3, 0.7\rangle \times 0.5 = \langle 0.5, 0.5\rangle.\end{aligned}$$

2. Updating this distribution with the evidence u_1 for $t = 1$ gives

$$\begin{aligned}\mathbf{P}(R_1 \mid u_1) &= \alpha\mathbf{P}(u_1 \mid R_1)\mathbf{P}(R_1) = \alpha\langle 0.9, 0.2\rangle\langle 0.5, 0.5\rangle \\ &= \alpha\langle 0.45, 0.1\rangle \qquad \approx \langle 0.818, 0.182\rangle.\end{aligned}$$

3. In case of repeated evidence, the probability of rain increases, since

$$\mathbf{P}(R_2 \mid u_1) \approx \langle 0.627, 0.373\rangle \quad \text{and}$$
$$\mathbf{P}(R_2 \mid u_1, u_2) = \alpha\langle 0.565, 0.075\rangle \approx \langle 0.883, 0.117\rangle.$$

Prediction

- **Prediction** is filtering without the addition of new evidence

$$\mathbf{P}(\mathbf{X}_{t+k} \mid \mathbf{e}_{1:t}) = \textstyle\sum_{\mathbf{x}_{t+k-1}}\mathbf{P}(\mathbf{X}_{t+k} \mid \mathbf{x}_{t+k-1})P(\mathbf{x}_{t+k-1} \mid \mathbf{e}_{1:t}).$$

where the parameter $k > 0$ (and hence $t + k - 1 \geq t$).

- The distribution $\mathbf{P}(\mathbf{X}_t \mid \mathbf{e}_{1:t})$ is obtained by filtering.

Example. Let us predict the chances for rain given u_1 and u_2:

$$\begin{aligned}\mathbf{P}(R_3 \mid u_1, u_2) &= \textstyle\sum_{r_2}\mathbf{P}(R_3 \mid r_2)P(r_2 \mid u_1, u_2) \\ &= \mathbf{P}(R_3 \mid r_2)P(r_2 \mid u_1, u_2) + \mathbf{P}(R_3 \mid \neg r_2)P(\neg r_2 \mid u_1, u_2) \\ &= \langle 0.7, 0.3\rangle \times p + \langle 0.3, 0.7\rangle \times (1 - p) \\ &= \langle 0.3 + 0.4p,\ 0.7 - 0.4p\rangle.\end{aligned}$$

where $\mathbf{P}(R_2 \mid u_1, u_2) = \langle p, 1 - p\rangle$.

Example (continued)

What happens if we make even further predictions into future?

1. For $k = 2$, we obtain

$$\begin{aligned}\mathbf{P}(R_4 \mid u_1, u_2) &= \textstyle\sum_{r_3}\mathbf{P}(R_4 \mid r_3)P(r_3 \mid u_1, u_2) \\ &= \mathbf{P}(R_4 \mid r_3)P(r_3 \mid u_1, u_2) + \mathbf{P}(R_4 \mid \neg r_3)P(\neg r_3 \mid u_1, u_2) \\ &= \langle 0.7, 0.3\rangle \times (0.3 + 0.4p) + \langle 0.3, 0.7\rangle \times (0.7 - 0.4p) \\ &= \langle 0.42 + 0.16p,\ 0.58 - 0.16p\rangle.\end{aligned}$$

2. In general, we have for any $k \geq 0$,

$$\mathbf{P}(R_{2+k} \mid u_1, u_2) = \langle 0.5 + (p - 0.5) \times 0.4^k,\ 0.5 + (1 - p - 0.5) \times 0.4^k\rangle$$

which converges towards the **stationary distribution** $\langle 0.5, 0.5\rangle$.

Smoothing

- The task is to compute $\mathbf{P}(\mathbf{X}_k \mid \mathbf{e}_{1:t})$ for $0 \leq k < t$ referring to past.
- Using a *backward message* $\mathbf{b}_{k+1:t} = \mathbf{P}(\mathbf{e}_{k+1:t} \mid \mathbf{X}_k)$, we obtain
$$\mathbf{P}(\mathbf{X}_k \mid \mathbf{e}_{1:t}) = \alpha \mathbf{f}_{1:k}\mathbf{b}_{k+1:t}.$$
- The backward message $\mathbf{b}_{k+1:t}$ can be computed using
$$\mathbf{b}_{k+1:t} = \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1} \mid \mathbf{x}_{k+1}) P(\mathbf{e}_{k+2:t} \mid \mathbf{x}_{k+1}) \mathbf{P}(\mathbf{x}_{k+1} \mid \mathbf{X}_k).$$
- Whenever $k+1=t$, the sequence $\mathbf{e}_{k+2:t}$ becomes empty and $P(\mathbf{e}_{k+2:t} \mid \mathbf{x}_{k+1}) = P(\top \mid \mathbf{x}_{k+1}) = 1$ where $\top$ stands for truth.
- This leads to a recursive definition, or algorithm
$$\mathbf{b}_{k+1:t} = \alpha \textsc{Backward}(\mathbf{b}_{k+2:t}, \mathbf{e}_{k+1:t}).$$

Example

Let us demonstrate smoothing with the umbrella example:

1. $\mathbf{P}(R_1 \mid u_1, u_2) = \alpha \mathbf{f}_{1:1}\mathbf{b}_{2:2} = \alpha \mathbf{P}(R_1 \mid u_1)\mathbf{P}(u_2 \mid R_1)$ where we already know the distribution $\mathbf{f}_{1:1} = \mathbf{P}(R_1 \mid u_1) = \langle 0.818, 0.182 \rangle$.
2. The distribution $\mathbf{b}_{2:2} = \mathbf{P}(u_2 \mid R_1) = \sum_{r_2} P(u_2 \mid r_2)\mathbf{P}(r_2 \mid R_1) = 0.9 \times \langle 0.7, 0.3 \rangle + 0.2 \times \langle 0.3, 0.7 \rangle = \langle 0.69, 0.41 \rangle$.
3. By substituting these distributions and normalizing, we obtain
$$\mathbf{P}(R_1 \mid u_1, u_2) = \alpha \langle 0.818, 0.182 \rangle \langle 0.69, 0.41 \rangle \approx \alpha \langle 0.564, 0.075 \rangle \approx \langle 0.883, 0.117 \rangle.$$
so that the smoothed estimate $P(r_1 \mid u_1, u_2) > P(r_1 \mid u_1)$.

☞ The additional piece of evidence u_2 increases the probability of rain on the first day, as the rain tends to persist.

Finding the Most Likely Sequence

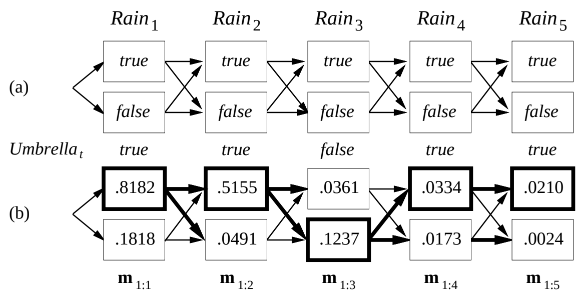
Example. Suppose that the security guard makes the following observations during the first five days: $u_1, u_2, \neg u_3, u_4, u_5$.

What is the weather sequence most likely to explain this?

- For each pair of states $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$, there is a recursive relationship between the most likely paths to $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$:
$$\begin{aligned} & \max_{\mathbf{x}_1 \ldots \mathbf{x}_t} \mathbf{P}(\mathbf{x}_1, \ldots, \mathbf{x}_t, \mathbf{X}_{t+1} \mid \mathbf{e}_{1:t+1}) \\ = \; & \alpha \mathbf{P}(\mathbf{e}_{t+1} \mid \mathbf{X}_{t+1}) \times \\ & \max_{\mathbf{x}_t} \left(\mathbf{P}(\mathbf{X}_{t+1} | \mathbf{x}_t) \max_{\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}} P(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t \mid \mathbf{e}_{1:t}) \right). \end{aligned}$$
- This equation is analogous to the one used in filtering.
- Maximization is performed for each value $\mathbf{x}_{t+1}$ of $\mathbf{X}_{t+1}$ in turn.

- In the filtering scheme, we have to replace $\mathbf{f}_{1:t} = \mathbf{P}(\mathbf{X}_t \mid \mathbf{e}_{1:t})$ by
$$\mathbf{m}_{1:t} = \max_{\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}} \mathbf{P}(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{X}_t \mid \mathbf{e}_{1:t})$$
and summation over $\mathbf{x}_t$ by maximization over $\mathbf{x}_t$.
- This gives the essential content of the **Viterbi algorithm** which has both linear time and space requirements.

Example. Consider the most likely explanation for $u_1, u_2, \neg u_3, u_4, u_5$:



3. HIDDEN MARKOV MODELS

- In a **hidden Markov Model**, or **HMM**, the world is described by a *single discrete* random variable X_t taking values $1,\ldots,S$ which correspond to the states of the world.
- The transition model $\mathbf{P}(X_t \mid X_{t-1})$ becomes an $S \times S$ matrix $\mathbf{T}$ such that $\mathbf{T}_{ij} = P(X_t = j \mid X_{t-1} = i)$.
- Forward and backward reasoning are simplified as follows:

$$\begin{aligned}\mathbf{f}_{1:t+1} &= \alpha \mathbf{O}_{t+1} \mathbf{T}^T \mathbf{f}_{1:t} \\ \mathbf{b}_{k+1:t} &= \alpha \mathbf{T} \mathbf{O}_{k+1} \mathbf{b}_{k+2:t}\end{aligned}$$

  where $\mathbf{O}_t$ is a diagonal matrix having $P(e_t \mid X_t = i)$ as the i^{th} value.
- For HMMs, the time and space complexities of forward-backward type reasoning are of the orders of $S^2 \times t$ and $S \times t$, respectively.

4. DYNAMIC BAYESIAN NETWORKS

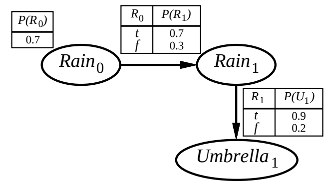
- A **dynamic Bayesian network** (DBN) represents how the state of the environment evolves over time.
- Each time slice of a DBN may have any number of state variables $\mathbf{X}_t$ and evidence variables $\mathbf{E}_t$.
- Every HMM can be transformed into a DBN and vice versa.
- By decomposing the state of a complex system into its constituent variables, the DBN is able to take advantage of the sparseness in the temporal probability model.

Example. The transition model of a DBN with 20 Boolean state variables, each of which has three parents in the preceding slide, has $20 \times 2^3 = 160$ probabilities while its HMM counterpart has 2^{40}.

Constructing Dynamic Bayesian Networks

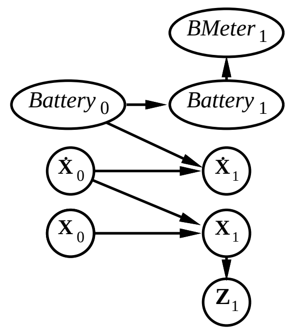
- To construct a DBN, one must specify three distributions: $\mathbf{P}(\mathbf{X}_0)$, the transition model $\mathbf{P}(\mathbf{X}_{t+1} \mid \mathbf{X}_t)$, and the sensor model $\mathbf{P}(\mathbf{E}_t \mid \mathbf{X}_t)$.
- For each time step t, there is one node for each state variable X_t and each evidence variable E_t plus relevant links between nodes.

Example. For the security guard example, it is sufficient to specify



Example

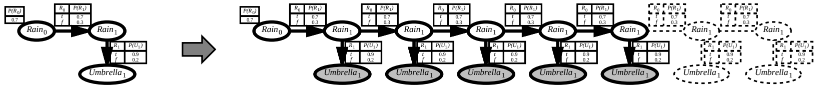
A robot is described with state variables $\mathbf{X}_t = \langle X_t, Y_t \rangle$ for position and $\dot{\mathbf{X}}_t = \langle \dot{X}_t, \dot{Y}_t \rangle$ for velocity and $Battery_t$ for actual battery charge level.



Both position (evidence variables $\mathbf{Z}_t$) and the battery charge level (evidence variable $BMeter_t$) are measured.

Exact Inference in DBNs

- The previous algorithms for inference in Bayesian networks can be applied to dynamic Bayesian networks.
- Given a sequence of observations, one can **unroll** a DBN until the network is large enough to accommodate the observations.
- Unrolling can also be done on a slice-by-slice basis.
- In the general case, the complexity of reasoning is exponential.



SUMMARY

- A dynamic world can be handled using a set of random variables to represent the state of the world at each point in time.
- Representations can be designed to satisfy the **Markov property**, so that the future is independent of the past given the present.
- With the **stationarity** assumption, i.e., the dynamics of the system does not change, much simpler probabilistic models are obtained.
- A temporal probability model consists of a **transition model** and a **sensor model**.